The double effect principle applied to ethical dilemmas of social robots

Bentzen, Martin Mose

Published in:
Proceedings - the Jin Yuelin Conference on Dao, Logic and Epistemology 2015

Publication date:
2015

Document Version
Peer reviewed version

The double effect principle applied to ethical dilemmas of social robots

Martin Mose Bentzen

Abstract

With the introduction of social robots into society arise new possibilities of cooperation between ethicists and roboticists, as social robots will be required to function in ways that are ethically acceptable, see e.g. (Wallach and Allen, 2009), (Arkin, 2009). However, ethics is traditionally a very informal field and robotics is a very technical one, so some problems of communication in the process of this cooperation are to be expected. Deontic logic can play a mediating role, at the same time providing tools for formalizing ethical principles in a way that can be justified from a philosophical point of view and to be used as a foundation for safe computational implementations. The overall project of devising a robot ethics can thus become a fruitful interdisciplinary field bringing insights together from philosophy, logic, and engineering, see e.g. (Wallach and Allen, 2009), (Bringsjord and Taylor, 2012).

In this paper, I take as starting point recent experiments with robots encountering ethical dilemmas reported by Winfeld and collaborators, see (Winfeld et al., 2014). I compare the dilemmas encountered by robots to some well-known thought experiments in ethics, see (Foot, 1967), (Thomson, 1985). I argue that a justifiable solution to the dilemmas will require that robots follow ethical principles which go beyond consequentialism. I first point to a number of possible ethical principles which may be implemented and discuss in some detail one such ethical principle, *the double effect principle*, which appeals to both consequentialist and deontological intuitions. The double effect principle states conditions for ethically acceptable behavior when there are both positive and negative consequences of an action. The action must itself be positive or neutral, the negative consequence may not be intended whereas the positive must be, the negative consequence may not be a means to obtain the positive effect, and the positive effect must be proportionally preferable to the negative effect, see e.g. (Mangan, 1949), (Foot, 1967), (Quinn, 1989), (McIntyre 2014).

The main contribution of the paper is a formalization of the double effect principle and a formal application of this principle to ethical dilemmas. In particular, I suggest how to handle intentions, causal reasoning and proportionality of several positive or negative consequences of an action in a given situation. I provide a formal semantics for the formalization which is based upon Action Type Deontic Logic, see (Bentzen, 2014). I conclude that although the double effect principle can be criticized from a philosophical point of view as guiding the actions of human beings, it is nevertheless fruitful to investigate implementations of the principle in robotics.

References

Arkin, R. (2009). *Governing Lethal Behavior in Autonomous Robots*, CRC Press.

Bentzen, M.M. (2014). Action Type Deontic Logic, *Journal of Logic, Language and Information*, **23**(4): 397-414, 2014.

Bringsjord, S. and Taylor, J. (2012). The divine-command approach to robot ethics, *in* P. Lina, K. Abney and G. A. Bekey (eds), *Robot Ethics: The Ethical and Social Implications of Robotics*, MIT Press: 85–108.

Foot, P. (1967). The problem of abortion and the doctrine of double effect, *Oxford Review* **5**: 5–15.

Mangan, J. (1949). An historical analysis of the principle of double effect, *Theological Studies* **10**: 41–61.

McIntyre, A. (2014). Doctrine of double effect, *in* E. N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, winter 2014 edn.

Thomson, J. J. (1985). The trolley problem, *The Yale Law Journal* **94**: 1395–1415.

Quinn, W. (1989). Actions, intentions, and consequences: The doctrine of double effect, *Philosophy and Public Affairs* **18**: 334–351.

Wallach, W. and Allen, C. (2009). *Moral Machines: Teaching Robots Right From Wrong*, Oxford University Press.

Winfield, A. F., Blum, C. and Liu, W. (2014). Towards an ethical robot: internal models, consequences and ethical action selection, *in* M. Mistry, A. Leonardis, M.Witkowski and C. Melhuish (eds), *Advances in Autonomous Robotics Systems*, Springer, pp. 85–96.